Vulnerabilities of the Modern Search Engine

Daniel Iuhasz
Computer Science and Automatics Faculty
"Politehnica" University of Timisoara
Timisoara, Romania
danieliuhasz@gmail.com

***Abstract*—The modern search engine is one of the most powerful and influential entities found on the internet. It controls the way traffic flows across the internet and to an account determines which online players lose or win money. Because of this, vulnerabilities of the Search Engine have been exploited with the intent of modifying the Search Engine Rankings to the advantage of some websites. A detailed view on under covered Vulnerabilities of the Modern Search Engine can be found in this article alongside a rundown of the most important techniques which take advantage of these vulnerabilities. This paper being a first in its domain will start with a thorough explanation of terms followed by a detailed view on each technique. To sum up, we will take a look at the ways search engines will evolve to cope with the ever-more present attacks and present further study directions which can be followed to details on certain topics.**

***Keywords*—search engine vulnerabilities, SEO, search engine optimization, search engine manipulation**

I. Introduction

A. *A Brief Introduction in the Anatomy of a Search Engine*

The basic search engine found in the 1900s used no solid metric for calculating the importance of a webpage – in other words, it was practically impossible to efficiently determine on a wide scale the importance on a single page on the World Wide Web. With the introduction of Google [1] the hyperlink was used as a way of quantifying quality.

Looking deeper into the Anatomy of the Search Engine, we find that the link is used as a vote. The more links a webpage has to it the more important it must be, to a specific domain. Using the Anchor Text, i.e. the visible text of the Hyperlink, as a domain or niche indicator enables the search engine to give accurate and good results based on the user's query. Consider a simple example: if everyone on the web is voting for one page when it comes to "pc repair", then that particular page must really be good at "pc repair".

Of course, the search engine has evolved a lot since its introduction, but the afore-mentioned explanation of the concept still stands. Nowadays, it has incorporated social integration, brand awareness filters and much more, but everything still works on the same principle, which makes it very vulnerable to attacks.

B. *Explanation of Terms*

This paper relies heavily on industry-specific terminology which is hard to understand without a proper explanation. Thus, this section is dedicated to explaining in short the terminology.

On-Page Search Engine Optimization is considered the optimization done on the page of the website. This is not actually a way to take advantage of the Vulnerabilities of the Modern Search Engine but is the base onto which all other techniques are built. The on-page SEO is done to make sure that a search engine can "read" the page and determine its relevant field, text, image attributes and other elements that relate to it.

Off-page SEO refers to everything which is done outside the webpage and includes, but is not limited to, linkbuilding, comment posting, social media interaction, microblogging. Everything that is done to indirectly raise the importance of a website is considered to be off-page SEO.

Black-hat SEO is a highly-discussed term in the industry. Some people relate it to every activity which is malicious, being on or off-page, but for the scope of this paper we will consider it to be only the on-page SEO that tries to fool the Search Engine into considering that a webpage has more content on it. This involves having one page for the search-engine and another for the human reader, which in other terms is regarded as a "back door".

Linkbuilding is the activity which has as a target the gathering of links which point to a certain website. The number and authority of links is important because they will be used by the Search Engine to compute the importance of a webpage. Links mimic the normal process of a page being "quoted" all over the internet and can be as simple as placing an article on another webpage or as complex as having a whole network of inter-linked sites to support linkbuilding.

Social Media Interaction is a technique which is based on social platforms for gathering links. This type of links are different from the normal ones that come from linkbuilding because they do not necessarily have a high value but they give the search engine signs that a subject is talked about on the internet. To make a clear distinction, we will have to consider an article talking about the Moon and a news item talking about a certain event that happened on the Moon. The first article will get more normal links and rank in the SERPS

far longer, because of its value, while the second article is just a newscast that has value for a number of days or months.

C. *Risks behind over-optimization*

Even if the techniques described in this article work both on a large and small scale, the webmasters who employ them have to take into account the fact that they are taking a high risk. Depending on how each particular search engine is implemented, there can be certain small or large penalties.

The most common penalty is the "temporary devaluation" of the website which causes it not to rank so well as before in the SERPS. This is harder to spot because of the natural database shift of the search engine, but in general if the website starts to rank less for a group of keywords or for all of them, then it is a "temporary devaluation".

A second penalty is referred to in slang as the "-950 penalty" and it refers to all keywords of a website ranking on the last page of the SERPS. Many times this penalty is temporary, but manual review is required to get a certain page/website back in the rankings. The "-950 penalty" usually is associated with a reduction of the metric used to calculate the importance of the whole website.

The last penalty is the total removal of the website from the index. This happens if the search engine has strong reasons to get the site deintexed and is associated with Black Hat techniques used to promote the website. In this case, under 0.1% of the websites that have been deindexed get back in the index.

In practice, we can see an example of such a penalty in the year 2006, when Search giant Google has "blacklisted" German car manufacturer BMW for breaching its guidelines [2].

II. On-Page SEO

When wanting to exploit the Vulnerabilities of a Modern Search engine, an Optimization Engineer will first ensure that a search engine can parse and understand that webpage. In this process, the engineer has to rely only on the structure of the HTML page to get the most out of a webpage. On the foundation of On-Page SEO, Black-Hat techniques can be applied to add additional information for the search engine to crawl.

A. *Domain Name*

The first indicative which makes a Search Engine determine the relevance of a webpage on a particular subject is the domain name. Choosing a relevant domain name which contains the main keywords a website wants to rank for is the best solution, but it could be a bad choice for brand ranking.

Inside the domain name all foreign characters are, at the time of this writing, considered by search engines a reason to discredit the trust of a website. To benefit from the vulnerabilities of a search engine, a domain name should not contain national characters or hyphens.

B. *Robots.txt and Sitepam.xml*

A website can block search engine access to it by placing in the Robots.txt special fields telling the web crawlers not to index the whole website or certain parts of it [3]. An example of a robots.txt file is the following:

User-agent: *

Disallow: /cgi-bin/

Disallow: /tmp/

Disallow: /~joe/

From experimental tests I have concluded that disallowing access to certain parts of a website that are not relevant will make the whole website rank better in the search engines. This is because the calculated rank of the whole website is split in the total number of pages.

An XML map of a site can be placed in the Sitemap.xml file which is usually found on the root of the server. The file should contain accurate and up-to-date information on the structure of the site, which will guarantee the fact that every new generated page is crawled as fast as possible.

C. *Title and Meta tags*

The title of a webpage is the most important indicative of what that page is about. By testing this I have concluded that no matter how much off-page optimization is done, the webpage will not rank as well without a proper title describing its contents. As a result, its impact is not only on the human reader but also on the search engine.

There are two types of meta tags which are important to search engines.

- The meta description tag gives a textual description of a webpage and is a short paragraph of text which should give an indication as to what the page is about. The syntax for it is: <meta name="description" content="Description Goes Here" />
- The meta robots tag gives a search engine information about whether the webmaster or the writer of the page wants this page to be indexed or not, or if the links of the page are followable. If the page has links which shouldn't be followed, then they are not used to compute the relevance of other pages in the search engine. The syntax for it is: <meta name="robots" content="values" />

D. *Frames, Embedded Players and JavaScript*

Every entity which cannot be well parsed by a search engine cannot take advantage of its vulnerabilities. Frames, Embedded Players and JavaScript links should in every case be avoided because they make the site lack the important content which is used to calculate the relevancy of a webpage. If the site does not have content, it will most likely not be trusted by a search engine [4].

E. *Links*

The last element which has to do with On-Page SEO is the link structure of a website. Every page on a website will have to have a link pointing to it from a more important page (also called an aggregator, a funnel or a category page). This way the power of a webpage can be split along multiple paths, raising the whole website and not just one of its pages.

To make a simple analogy, the power of a website can be compared to water flowing along a series of canals to individual lakes. Each small lake is a terminus webpage, which has to be connected to the water source – the main page or the root. If there are no links from the root to the lake, there will be no water in the lake.

III. Off-Page SEO

A. *Linkbuilding*

The linkbuilding process can be described as the technique used to manually or automatically acquire links from webpages on the web, in an unnatural way. By doing so, the owner of the webpage makes it seem like his/her particular price of writing is actually relevant on a certain domain.

When doing linkbuilding, the author must place a link destination address and an anchor text. The anchor text is the visible text of the link and many search engines consider it to be a sign of relevancy. For instance, if linking with the text "black car" to a webpage, then that page is most likely about black cars.

By building a lot of links towards a webpage, the search engine is fooled into ranking it for a certain keyword. The time it takes for a webpage to rank is dependent on the domain – for some pages it could take just a few links while for others it might never work. A proper analysis is done beforehand to research which keywords are worth pursuing, but in general with enough links a page can rank for anything.

Another important technique in SEO is the use of silo link farms (Fig.1) to deliver very powerful links towards a webpage. Link trust is being exponentially increased by adding an auxiliary page between the link and the webpage it should link to. In simple linkbuilding links are being built from page A to page B (A->B). In silo link farms SEO engineers build links from page A to page C and from page C to page B (A->C->B). As a result, building more links to page C will also add more power to page B and also minimize the change of being spotted by automated linkbuilding filters implemented by search engines.

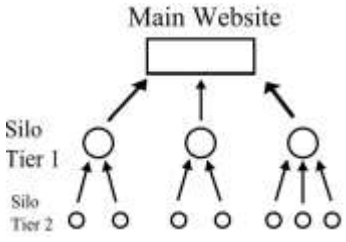

Figure 1. Silo Linkbuilding

B. *Comment Posting*

Comment Posting is similar to Linkbuilding but differs in the sense that it gives a website the basic Social Media Interaction it needs in order to become more brand-oriented. Search Engines usually trust brand more than lesser-known websites, so it is normal to try to become a brand. The name of the brand is not important, but if many people are commenting on blogs with a normal name as an anchor text, then that particular website is considered "brandable".

In comment posting the links are not targeted to the internal pages of a website, but only to the homepage. This means that a proper link structure has to be in place for the link power to flow from the homepage to all other pages of the website, but the trust comment posting brings will double the normal linkbuilding in order to produce a variety of links.

It is expected that near the end of the year 2011, comments will no longer carry that much power as they used to. This doesn't mean that they are not important for a diversity of links – creating a natural link profile will minimize the risks of taking advantage of the Vulnerabilities of Modern Search Engines.

C. *Social Media Interaction*

The newest and most controversial method of building website authority in search engines is Social Media Interaction. Modern Web 2.0 platforms make it very easy to share content on the web and enable users to link to a multitude of sites. Links can be acquired from such sources, but they rarely carry any power when it comes to ranking a website. This is because some domains are not usually linkable (such as certain parts of medicine, for instance), and entertainment sources are usually the ones which get a lot of social media attention.

From experimental tests on one website with Web 2.0 linking and another with no Web 2.0 interaction, I have concluded that the modern search engine computes trust as a value of the number of users which come from a trustworthy source. This means that if a website brings a lot of traffic from a good website, it is worth trusting.

With a lot of automated plug-ins that enable webmasters to post a link on Web 2.0 platforms for every page of a website, it is important to have a good following which goes to that particular page.

IV. Conclusion

Search Engines, as every piece of software rely on invariants to give accurate results to a search query. By modifying these invariants the average website can rank well in search engine results and are able to compete with large websites. This gives every small team the ability to build a steady and on-going resource online that would otherwise not be able to surface.

The fragility of the system, however, makes using the Vulnerability of Modern Search Engines as a business model a bad choice. A website will have to offer values in order to depend only partially on its online rankings.

Acknowledgment

This paper is written as a part of a continuous learning process in the less-known field of Online Website Optimization. While normal programming techniques target the fulfillment of stakeholder requirements such as website speed, security of style, OWO aims at bringing business-critical traffic to a website.

I hope that in the future more people will understand the importance of optimization and not consider this area as just a subfield of Online Marketing.

References

[1] S. Brin and L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine", Stanford University, http://infolab.stanford.edu/~backrub/google.html

[2] BBC News, 6 February 2006, http://news.bbc.co.uk/2/hi/4685750.stm

[3] D. Dover, "Search Engine Optimization (SEO) Secrets", Wiley Publishing, Inc., Indianapolis, 2011

[4] E. Enge, S. Spencer, R. Fishkin, and J. C. Stricchiola, "The Art of SEO", O'Reilly Media, Inc., 2009